D652–D657 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
doi:10.1093/nar/gkp995

COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer

Simon A. Forbes, Gurpreet Tang, Nidhi Bindal, Sally Bamford, Elisabeth Dawson, Charlotte Cole, Chai Yin Kok, Mingming Jia, Rebecca Ewing, Andrew Menzies, Jon W. Teague, Michael R. Stratton and P. Andrew Futreal*

Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Received September 15, 2009; Revised October 15, 2009; Accepted October 16, 2009

ABSTRACT

The catalogue of Somatic Mutations in Cancer (COSMIC) (http://www.sanger.ac.uk/cosmic/) is the largest public resource for information on somatically acquired mutations in human cancer and is available freely without restrictions. Currently (v43, August 2009), COSMIC contains details of 1.5-million experiments performed through 13 423 genes in almost 370 000 tumours, describing over 90 000 individual mutations. Data are gathered from two sources, publications in the scientific literature, (v43 contains 7797 curated articles) and the full output of the genome-wide screens from the Cancer Genome Project (CGP) at the Sanger Institute, UK. Most of the world's literature on point mutations in human cancer has now been curated into COSMIC and while this is continually updated, a greater emphasis on curating fusion gene mutations is driving the expansion of this information; over 2700 fusion gene mutations are now described. Whole-genome sequencing screens are now identifying large numbers of genomic rearrangements in cancer and COSMIC is now displaying details of these analyses also. Examination of COSMIC's data is primarily web-driven, focused on providing mutation range and frequency statistics based upon a choice of gene and/or cancer phenotype. Graphical views provide easily interpretable summaries of large quantities of data, and export functions can provide precise details of user-selected data.

INTRODUCTION

The Catalogue of Somatic Mutations in Cancer (COSMIC) database is designed to annotate, hold, navigate and display the published literature on somatic mutations in human cancer. Huge amounts of information have been generated on this subject, but these are difficult to examine collectively due to their format or distribution. COSMIC collects all this detailed information together in one place by precisely and exhaustively curating the scientific literature, both for known cancer genes and large systematic screens, and combines it with the large analyses from the Cancer Genome Project (CGP) at the Wellcome Trust Sanger Institute, UK. The website over this database is designed to provide an easy interface to query this complex data, providing graphical and tabulated aggregate views allowing detailed examinations and summary analysis of mutation range and prevalence, both from a gene and cancer type perspective. Most recently, whole-genome sequencing results have been included, overviewed in a new circular diagram with separate displays for navigation and viewing of this new data.

DATABASE CONTENT

The data in COSMIC is derived from two sources. The literature curation effort uses original publications as the data source to maximize the precision of the data obtained, no indirect curation of external databases has been performed. Genes are highlighted for curation by their presence in the Cancer Gene Census [(1); http://www.sanger.ac.uk/genetics/CGP/Census/], focusing on these well-characterized genes and collecting large quantities of information (for example, over 13 000

*To whom correspondence should be addressed. Tel: +122 349 4730; Email: paf@sanger.ac.uk

instances of 105 unique sequence variants have been curated in KRAS). Genes promoting cancer via point mutations have been primarily targeted since they have a great impact in carcinogenesis and a large quantity of information to confidently generate mutation range and prevalence statistics. Cancer-promoting gene fusion mutations are also being curated, as there are many more genes promoting tumour growth via fusion events than point mutations (1). The literature curation process is largely manual, gradually accumulating the quantity of data currently available over 9 years. As information is entered to COSMIC by dedicated trained curators, software systems check each datapoint for accuracy and integrity. Large systematic candidate gene screens are additionally being curated in a semi-automated fashion, with initial interpretation of large result sets automated, followed by further manual and software checks. This should allow the inclusion of all somatic mutations with potential impact in human cancer. These will serve as a resource for whole cancer genome resequencing data. All literature curation genes in COSMIC are updated each release (six releases are scheduled per year) with data curated from recent papers to present the most recent somatic mutation reports.

Additionally, COSMIC serves as the portal to access in-house somatic mutation screening data from the systematic resequencing studies undertaken by the CGP. Such information is now available for over 4000 genes screened. More recently, CGP data includes low coverage whole genome resequencing screens (2), utilizing next-generation sequencing technologies which can precisely identify genome-level rearrangements previously invisible in both karyotypic and candidate gene screen procedures. Each mutation found has been resequenced several times and examined by scientific staff for confident confirmation.

All annotations in COSMIC require the tumour sample to have a phenotypic classification. This comprises a description of where the tumour was found (e.g. lung) and its morphology (e.g. non-small cell lung cancer, NSCLC). Usually broken down into increasingly specialized definitions (e.g. skin, face, NS: malignant melanoma, lentigo maligna, NS), these allow the mutation data to be navigated at different levels. For instance it is useful to examine KRAS mutations from a broad primary site perspective; lung tumours show a very different mutation spectrum in this gene than large intestine cancers. However, examining Lentigo Maligna (a subclassification of melanoma) in isolation, shows mutations in only four genes, only two of which have significant frequencies (BRAF, NRAS). COSMIC's phenotype definition system has been built for this purpose, allowing layered and very precise classification of each tumour. Rather than adhere to an existing standard of tumour classification, COSMIC has developed it from the ground up to be as useful as possible for this purpose. During curation the published classification for each tumour is recorded before being translated into COSMIC's system. This system (not independently published) is described in further detail on COSMIC's 'additional info.' page (http://www.sanger.ac.uk/genetics/CGP/cosmic/add_info/), including a listing of classification translations.

Each mutation recorded in COSMIC is given a unique identifier, a series of details and a single statement that precisely defines it (e.g. 'c.1799T>A'; 'p.V600E'; 'chr2:g.140099605T>A'). This mutation 'syntax' is derived from the HGVS nomenclature recommendations (3). Coding domain annotations receive a 'c.' syntax, indicating the mutation is located on a cDNA CDS co-ordinate and a 'p.' syntax indicating an amino acid co-ordinate. Gene fusions receive an 'r.' syntax indicating the co-ordinate is relative to the gene pair's mRNA co-ordinates, an important distinction as the UTR sequences can be significant in these novel products. Finally, mutations can also receive a 'g.' syntax indicating nucleotides on the golden path which are involved in the mutation. Details of these syntaxes are discussed at length at http://www.genomic.unimelb.edu.au/mdi/mutnomen/.

DATA ACCESS

The website is focused on presenting complex phenotype-specific mutation data in an easily interpretable format, usually best presented graphically. Initial entry into the system is usually via selection of a gene or cancer tissue type (phenotype), either using 'browse by' features where one is presented with a list to choose from, or the 'search' box where any text can be typed for a COSMIC database search. Whilst browsing by gene is fairly straightforward, browsing by phenotype can become complex, with multiple increasingly specialised phenotype choices; for instance one can navigate to 'lung' only, or to the very specific 'lung, right lower lobe; carcinoma, adenocarcinoma'. Acceptable search terms include not only gene names or phenotype terms, but also COSMIC database IDs, or mutation description syntaxes (e.g. 'p.V600E'). Once a choice is made, summary information is presented with mutation counts and frequencies; the gene summary page provides a mutation spectrum map and many links to external resources; the phenotype (tissue) summary page provides lists of mutated genes. These pages lead into the histogram page, where most examination of the data occurs. Mutations are graphically displayed for the selected gene, as a histogram of single-base substitutions, together with a map of insertions, deletions and complex replacement mutations (Figure 1). The histogram immediately shows the type of cancer gene selected, easily identifiable by its mutation signature. Figure 1a shows BRAF, a gain-of-function oncogene, requiring very specific sequence changes for oncogenic activation [the missense mutation p.V600E represents 75% of all muations found in this gene (n mutated = 13 885)]. Conversely, Figure 1b shows CDKN2A; cancer can only progress after this key tumour suppressor gene is inactivated. As the histogram shows, this can happen in a number of ways across the entire gene's length (n mutated = 3262). Tabulated breakdowns of the data are available under the graphic either by phenotype or sequence change, and additional messages will appear if further fusion mutation data is available for the gene. The histogram can display mutations either by CDS or amino acid co-ordinates and, most usefully, can be zoomed to

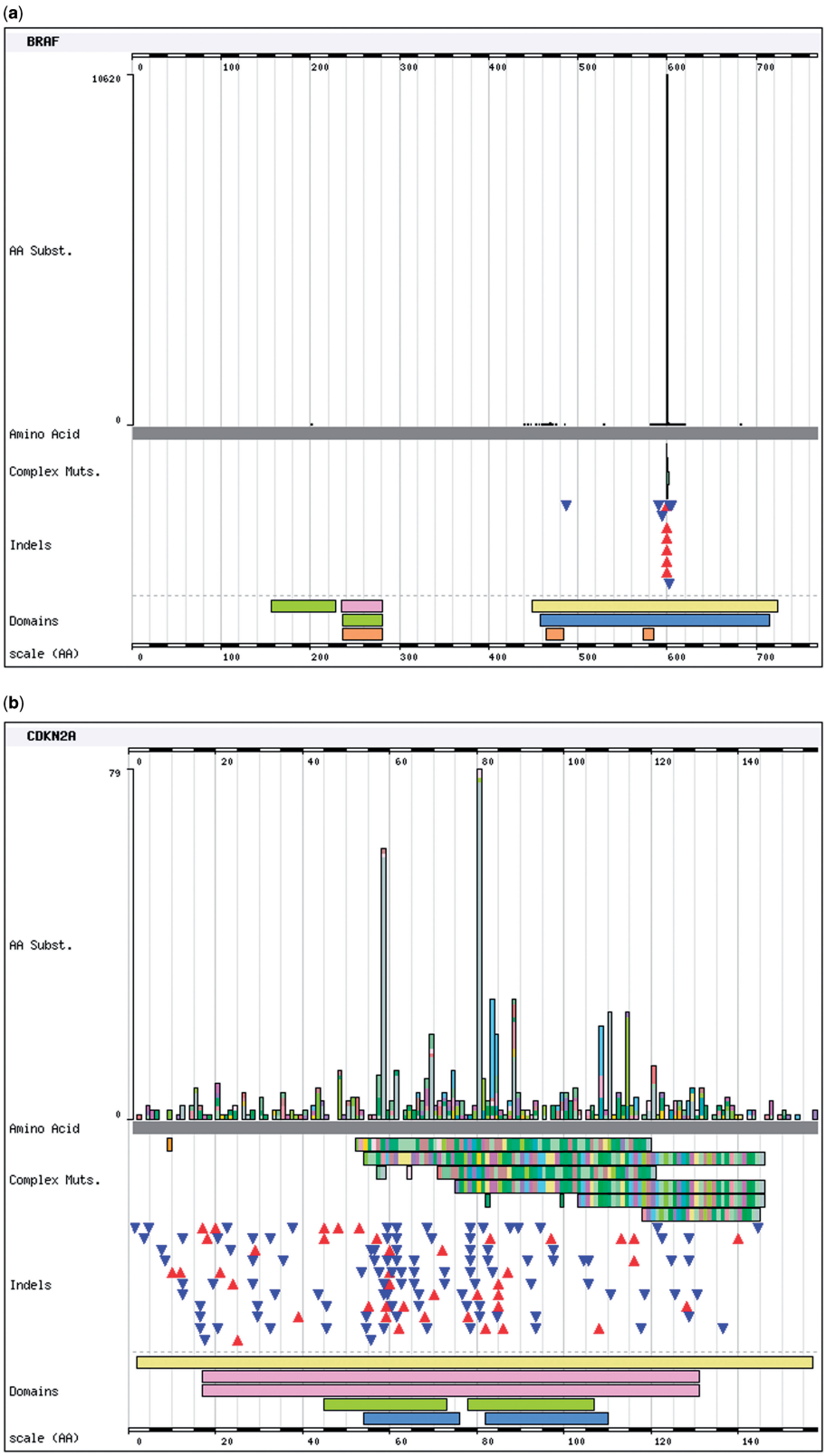

Figure 1. The Histogram page forms the core of the COSMIC website, showing the mutation range for each gene selected. The frequencies of each single base substitution are shown in the histogram itself and each deletion, insertion and complex substitution are indicated below as triangles (deletions point down in blue, insertions point up in red) or bars (which indicate the length of sequence being replaced). Below the mutations, long single-colour bars indicate structural domains, providing links to external databases including Pfam and InterPro. Two histograms are shown to demonstrate the range of mutation distributions observed; (**a**) BRAF shows the classical gain-of-function spectrum, whereby very specific mutations are required to activate growth promotion, in this case most frequently a c.1799T>A substitution; (**b**) CDKN2A (p16) is a tumour suppressor gene, only allowing growth promotion when its activity is absent ('loss-of-function'). Tumourigenic mutations only need to inactivate the gene, shown here as a range of mutations across the gene's length, particularly including a large quantity of frame shifting insertions and deletions.

scrutinize particular regions of a gene more closely, with the tables below reflecting the details in the zoomed region of the graphic. Additionally, once a gene is being viewed, a phenotype can be specified (at any level of detail) and both the graphic and table will be redrawn to reflect the new parameters. In this way, the histogram page can be used in a process of circular navigation, constantly specializing from general to specific queries, gaining much knowledge in the process. Links from this core page direct users to summaries and further details of mutations and publications, together with an export function, to output the information onscreen in a spreadsheet format.

Gene fusion mutations are indicated within the summary and histogram pages, but in separate sections; if a gene is fused with another, or a sample contains such a mutation, a separate indicator is shown to direct the user to a summary page which overviews the mutation frequencies and phenotypes associated with fusions between the two selected genes, again with graphic and tabulated displays. Fusions are defined by exon content, including UTR exons which are significant in some fusion events, leading to displays largely focused on defining the breakpoint with respect to the nearest exon of each gene. The 'r.' mutation syntax records not the breakpoint position, but the mRNA content from each gene (including intronic or non-templated sequences were appropriate).

Samples which have undergone whole-genome sequencing receive summary annotation in the form of a 'circos' diagram [Figure 2; (4)]. This is due to the quantity and complexity of the data collected by these studies, which includes extensive lists of coding and non-coding point mutations, copy number variations and genomic rearrangements. While most of this information can be displayed in linear tracks against standard karyotype ideograms, the non-linear nature of interchromosomal rearrangements is much more suited to a circular diagram, and COSMIC has now standardized on this format for whole-genome sequencing experiments. All known genotypic information for a sample is included in the diagram, regardless of its originating methodology. Forming a starting point for the investigation of these highly annotated samples, the ring of coding mutations provides links into the standard cosmic gene-centric views, the CNV track links into the CGP's SNP array-based CGH site (http://www.sanger.ac.uk/cgi-bin/genetics/CGP/CGH_home.cgi) for further evaluation of copy number abnormalities, and the structural rearrangements link into a subsystem of pages to examine these in more detail.

The main COSMIC website (blue pages) overviews the whole database, showing all curated information. However, within the larger COSMIC websystem, multiple subdomains exist to navigate subprojects independently, each colour-coded for easy recognition. A yellow page (http://www.sanger.ac.uk/genetics/CGP/Classic/) summarizes the information curated from the scientific literature providing summaries and links into COSMIC. Two entire websites, colour-coded copies of the main cosmic system, overview the CGP's contribution to COSMIC. The CGP resequencing studies, focused candidate gene screens sequencing large numbers of tumour samples through over 4000 genes are presented on a red COSMIC site (http://www.sanger.ac.uk/genetics/CGP/Studies/). In green, the cancer cell line project displays results of a systematic characterization of the mutant genotypes of almost 800 well-known cancer cell lines (http://www.sanger.ac.uk/genetics/CGP/CellLines/). Particularly in the green website, mutations are only included if they are known, or are highly likely to, have a significant impact on the oncogenesis of the tumour examined; an additional datasheet is available describing mutations of unknown significance resulting from the cell line project.

While the COSMIC websites are geared toward easy graphical navigation, custom data mining and large-scale analysis often requires the data in a different format. For this purpose, the content of each release is summarized in a series of very large spreadsheets and placed on a dedicated FTP site (ftp://ftp.sanger.ac.uk/pub/CGP/cosmic) for downloading. For those needing to mine the full content of COSMIC, each release is also available as a relational database export in Oracle 10 format, but supporting this requires significant IT infrastructure.

COSMIC data is also available outside of the COSMIC system. Where possible, all co-ordinate data in COSMIC has been located on the human genome (i.e. genes, exons, mutations). This has allowed the integration of COSMIC data into the Ensembl genome browser (http://www.ensembl.org) via DAS [distributed annotation system, (5)]. Once COSMIC DAS tracks are turned on (either via Ensembl configuration or a link on the gene or mutation summary pages), COSMIC annotations for both genes and mutations can be examined in a genomic context (Figure 3). For the 62 main curated genes, mutation annotations in a peptide structure context (uniprot annotations) are also available as a DAS track. COSMIC DAS tracks are all available through http://www.dasregistry.org/. A much lengthier discussion of the data in COSMIC and its navigation is available as a Protocol document (6).

FUTURE WORK

As the complexity and scope of the data in COSMIC expands, more tools are needed to mine the data and reference it to external systems, and new systems are currently being built with this in mind. Especially notable amongst these upgrades is a new Biomart which will soon be available for COSMIC. Especially designed for the mining of complex datasets, Biomart (7) allows the specification of any number of predicate sets, querying the COSMIC database and returning the appropriate data in tabulated form, with links back in to the COSMIC websystem where appropriate. This has been designed to encompass the full scope of the published COSMIC data rather than any subdomain. One of the strengths of Biomart is its ability to federalize several independent systems and query them together returning a coherent dataset combined from all. Using the external database identifiers in COSMIC already, the COSMIC biomart

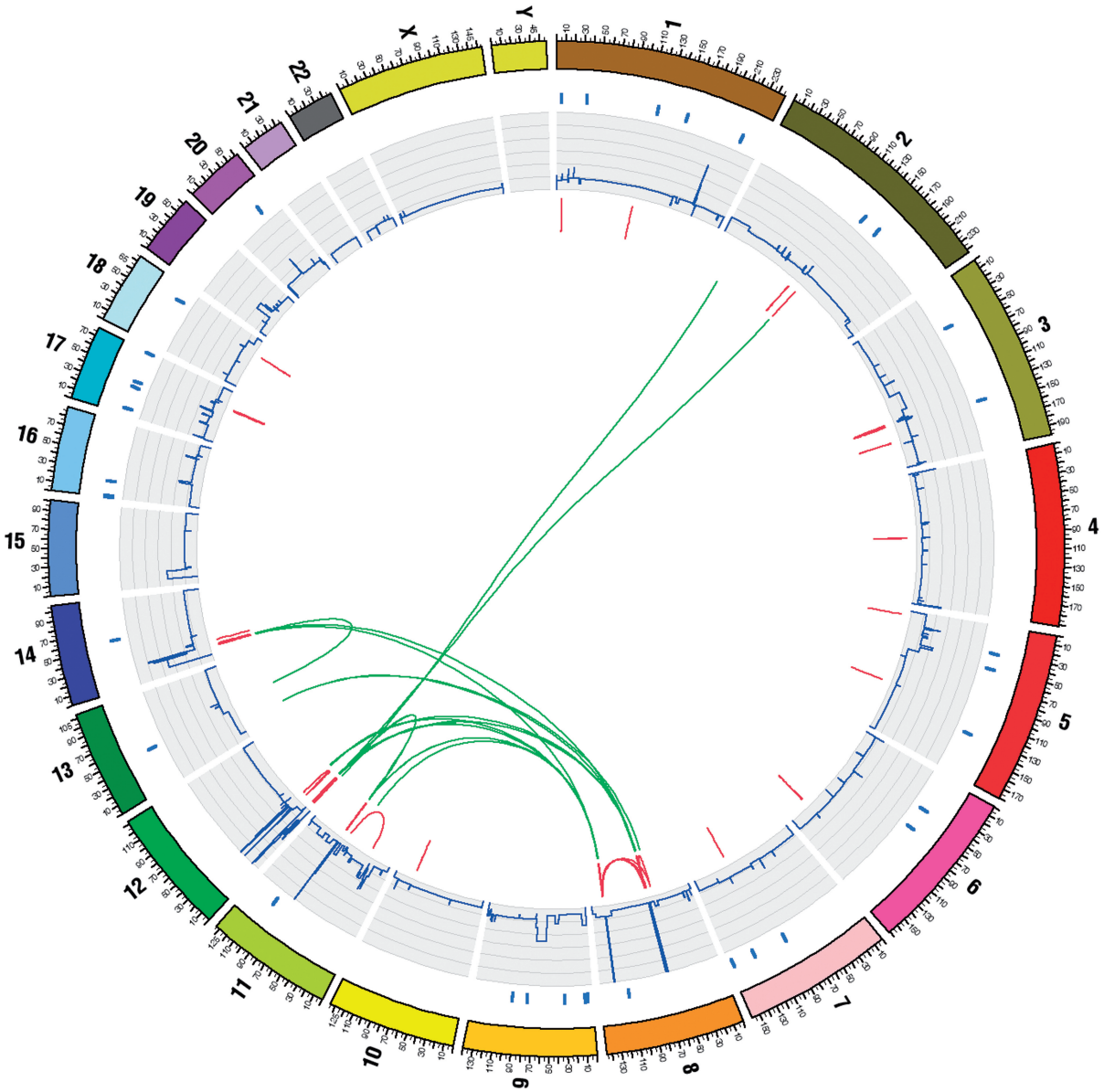

Figure 2. COSMIC has now standardized on Circos diagrams (4) to display genome-wide mutation analyses for a sample. Concentric rings display different data for the same sample, all located to the same genomic co-ordinates. On the outside of the circle, the chromosomes of the standard reference genome are coloured, numbered and appropriately scaled. Moving from the chromosome indicators to the centre of the circle, blue lines indicate the presence of coding point mutations in the coding domains of genes, a ring of histograms indicate the copy number of each genomic segment, red bars indicate the presence of small intra chromosomal rearrangements and green lines link chromosomes where large intra chromosomal rearrangements have been observed.

has been federated to Ensembl to allow the retrieval of gene and genome annotations surrounding regions of interest in cancer. Further federalization is possible and we are examining other ways of integrating COSMIC with external resources. It is expected this addition to COSMIC should be available before the end of 2009.

DISCUSSION

COSMIC comprises a number of subtly different data domains. Ninety percent of the mutation content adheres to the original goal of curating the world's literature on known cancer genes. However, with the speed of DNA sequencing accelerate rapidly, tumour analysis is becoming less focused on single genes and it is not unusual to find studies which instead analyse thousands. COSMIC has held CGP data from these screens for years due to its use as a prepublication data release site for primary tumour and cell line analyses. However, curating published systematic screens has traditionally been more difficult due to the large number of mutations and genes, and the lack of automatable informatic access. In order to include these large-scale studies in COSMIC, the curation systems have now been extended and the inclusion of such data has begun with the release of the Sjoblom *et al.* screen of breast and colorectal tumours in

Figure 3. COSMIC mutations are available in the Ensembl genome browser, into which all other genomic annotations can be combined. Each unique sequence change is separately indicated and in a genomic context, the coding bias of COSMIC's data makes the annotations pile up over the exons. In this example, the first 15 rows of mutations in the PTEN gene is shown (reduced from 197 rows in Ensembl).

the full CCDS gene set (8). While the curation of point mutations in COSMIC is an ongoing activity, the number of uncurated genes remaining is rapidly reducing and has allowed a redirection of effort to include oncogenic gene fusion events, with many more genes involved than point mutations (1). The genetic analysis of human cancer has also begun to lose its coding-domain focus, with full-genome resequencing of tumours now underway, generating tens of thousands of mutations per sample, together with many genomic rearrangements. These are now being released into COSMIC, using the Circos diagram (4) as a basis for summarizing genome-wide data, particularly useful for presenting complex intra- and inter-chromsosmal relationships. COSMIC will continue to curate the world's literature on somatic mutations in cancer, but it will increasingly develop with the aim of displaying and navigating data from whole cancer genome resequencing efforts.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Wellcome Trust for support under grant reference 077012/Z/05/Z.

FUNDING

Funding for open access charges: Wellcome Trust (grant reference 077012/Z/05/Z).

Conflict of interest statement. None declared.

REFERENCES

1. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
2. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
3. den Dunnen,J.T. and Antonarakis,S.E. (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mut.*, **15**, 7–12.
4. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S. and Marra,M. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
5. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.
6. Forbes,S.A., Bhamra,G., Bamford,S., Dawson,E., Kok,C., Clements,J., Menzies,A., Teague,J.W., Futreal,P.A. and Stratton,M.R. (2008) The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.*, **10**, 11.
7. Haider,S., Ballester,B., Smedley,D., Zhang,J., Rice,P. and Kasprzyk,A. (2009) Biomart Central Portal - unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.
8. Sjoblom,T., Jones,S., Wood,L.D., Parsons,D.W., Lin,J., Barber,T.D., Mandelker,D., Leary,R.J., Ptak,J., Silliman,N. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.